94-06

Statistical issues raised by the Bellcore data

N.G. Duffield[1,2], J.T. Lewis[2], Neil O'Connell [2,3],
Raymond Russell [2] and Fergal Toomey[2]

Introduction There has been a recent surge of literature claiming that "Ethernet traffic is self-similar" and possesses long range dependence [12, 13, 14]; similar claims have been made with reference to other forms of telecommunications traffic [2]. Leland *et al.* [12, 13, 14] have justified these claims using traffic observations that were taken from an Ethernet local area network at Bellcore Laboratories: we refer to these observations, which have been made publicly available, as the 'Bellcore data'. In this lecture we discuss the implications of such claims for the problem of estimating loss probabilities in networks, and their validity.

There has been much recent work on estimating rare event probabilities in queueing networks, in as much generality as possible [1, 3, 9, 7, 8, 10, 11]. This has been motivated by potential applications in the design and performance of high-speed telecommunications networks.

The starting point in the development of a general network theory is to consider what happens in a general single server queue. Roughly speaking, for a stable queue with deterministic service rate, under very general conditions on the arrivals process, the tails of the distribution of queue length Q should satisfy

$$P(Q > b) \approx e^{-\delta b}, \tag{1}$$

for some positive constant δ that depends on the service rate at the queue and the statistical properties of the arrivals process. A more precise statement of this fact can be found in the appendix. Roughly speaking, a sufficient condition for (1) to hold is that the arrivals process is stationary and mixing (ie. does not possess long range dependence).

It is thus a very general result, suggesting potential applications to real traffic problems. For example, it can be used for estimating overflow probabilities in very large buffers, using observations of traffic over relatively short time periods. This idea is originally due to Courcoubetis *et al.* [4]; they propose a method, justified by (1), of extrapolating from the observed tail frequencies of the queue. We have proposed a different method [6] which also relies on the validity of (1).

The first question we address is: *should we expect (1) to hold for self-similar traffic that possesses long range dependence?* The answer is 'no': this follows from a theorem in [7]. In fact, self-similarity is not an issue here: it is the presence of long range dependence that destroys the property (1). We refer the reader to the appendix for more details.

We have therefore been encouraged to take these claims of long range dependence quite seriously and question their validity.

Inferring long range dependence. Long range dependence is often used as a a possible explanation for unexplained variation at every time scale over which a process is observed. However, if the variation at the larger time scales can be explained by additional information, the question of long range dependence may become redundant. For example, consider a sequence of independent random variables $X_1, \ldots, X_{2n}$: $X_1, \ldots, X_n$ have a normal distribution

[1]School of Mathematical Sciences, Dublin City University, Dublin 9, Ireland
[2]Dublin Institute for Advanced Studies 10 Burlington Road, Dublin, Ireland
[3]Author presenting paper at the 11th UK Teletraffic Symposium

with mean 20 and unit variance and $X_{n+1}, \ldots, X_{2n}$ have a normal distribution with mean 10 and unit variance. Suppose we were to observe a realisation of this process, in ignorance of how it was generated. The sequence will almost certainly pass any statistical test for long range dependence that aims to detect slow decay rates in autocorrelation or, equivalently, non-linear growth in variance through aggregation. When we look at the data we will almost certainly see a 'level shift' about half way along the sequence.

What should we think? That the observed sequence is embedded in a longer sequence that exhibits long range dependence is a possible explanation: the level shift is a random fluctuation at the order n time scale, and there are similar random fluctuations at every time scale... It is also a rather fanciful explanation, and certainly not very useful. A more cautious response would be to simply accept that there is a level shift in the data and that the process is displaying two distinct types of behaviour, depending on whether it is observed before or after the shift; a closer inspection would suggest that the observations are independent given the location of the shift. Either way, we are not really in a position to predict future behaviour; not without access to more information.

We could investigate the source of the data. Suppose, for the sake of argument, that we are told the data represents measurements of Ethernet traffic levels taken from a local area network in a a Government office between the hours of 4pm and 6pm on a Friday evening; on further inquiry we discover that half the employees leave the office at 5pm sharp while those remaining work until 6pm. Long range dependence seems suddenly implausible. The level shift we observed in this data is indeed a fluctuation that occurs on a much larger time scale than the length of the observation period, but it is not random, or even 'random': it is a *periodic* event, occurring once a week.

These remarks are quite relevant to our discussion of the Bellcore data.

The Bellcore data. Leland *et al.* [12, 13, 14] have analysed Ethernet traffic measurements taken from local area networks at Bellcore Morristown Research and Engineering Center between August 1989 and February 1992; they claim that the traffic is statistically self-similar and exhibits long range dependence. A variety of statistical tests were performed, including the inspection of 'variance-time plots'. We will concentrate on the use of variance-time plots as a detector of long range dependence, as the rationale behind this approach is easily explained and it is essentially equivalent to the other tests that were performed.

Denote by $X_1, \ldots, X_n$ the traffic measurements over a period of observation. A variance-time plot can be produced as follows. We begin by computing, for each m, an aggregated sequence

$$X_k^{(m)} := \frac{1}{m} \sum_{j=(k-1)m+1}^{km} X_j, \qquad k = 1, 2, 3, \ldots \tag{2}$$

Then for each m, we compute the sample variance v_m of the sequence $X_1^{(m)}, X_2^{(m)}, \ldots$, and plot $\log v_m$ against $\log m$. If the observations are taken from a stationary sequence that does not exhibit long range dependence we expect to observe, given sufficient data, an asymptotic slope of -1; on the other hand, for a stationary sequence with long range dependence, we expect the observed slope to be strictly greater than -1 for large values of m (a self-similar process with Hurst parameter H will produce an asymptotic slope of $-2(1-H)$).

Leland *et al.* [12, 13, 14] consistently observed slopes that were greater than -1. The variance-time plots for two of their data sets are shown in Figure 1. One of these data sets is about

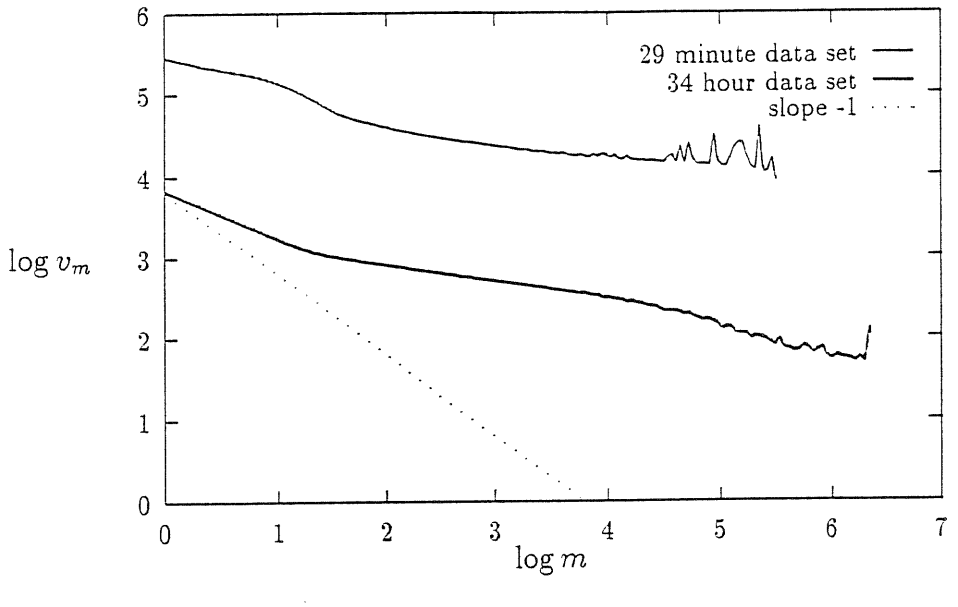

Figure 1: Variance-time plots for two of the Bellcore data sets.

30 hours in length; the other is only 29 minutes long. If we were to conclude that the traffic is self-similar with Hurst parameters greater than 1/2, we would predict (see appendix) that the log-frequency of overflow observed when this traffic is fed through a queue with constant service rate behaves like $-b^{2(1-H)}$ for large b; a value of $H > 1/2$ would therefore have a major influence on our predictions.

Let us now have a closer look at the 29-minute data set. We have aggregated this data and recorded the number of bytes observed in each 10-second time interval: the outcome is displayed in Figure 2. There is clearly a level-shift here, and the regions labelled II and IV seem quite stationary. In fact the observations within each of these periods appear to be almost independent. We have thus created variance-time plots for each of these regions (Figure 3) and, not surprisingly, the slopes at large aggregation levels are quite close to -1. We have no idea why there is a level shift in this data set, but it is certainly not evidence for long range dependence.

The 30-hour data set is displayed in Figure 4 using aggregation levels of 240 seconds. Clearly, there are varying mean levels of activity on this time-scale. The simplest explanation for this is that the time of day is an important factor in determining levels of activity. Note that 'time of day' is a periodic process. This simple observation renders the claims of long range dependence to be unfounded: it is important to remove all periodicities in the data before performing tests for long range dependence, as these will distort the statistics if there is not enough data on the time-scale at which these periodicities occur (in other words, there is not enough data here to remove the time-of-day effect).

Final remarks We conclude this lecture with some remarks on the use of the formula (1) in practical situations. First, if a traffic stream is stationary over time periods which are long compared with the range of dependence, then (1) may be useful for short term prediction and hence dynamic allocation of resources. The range of dependence will depend on the number of

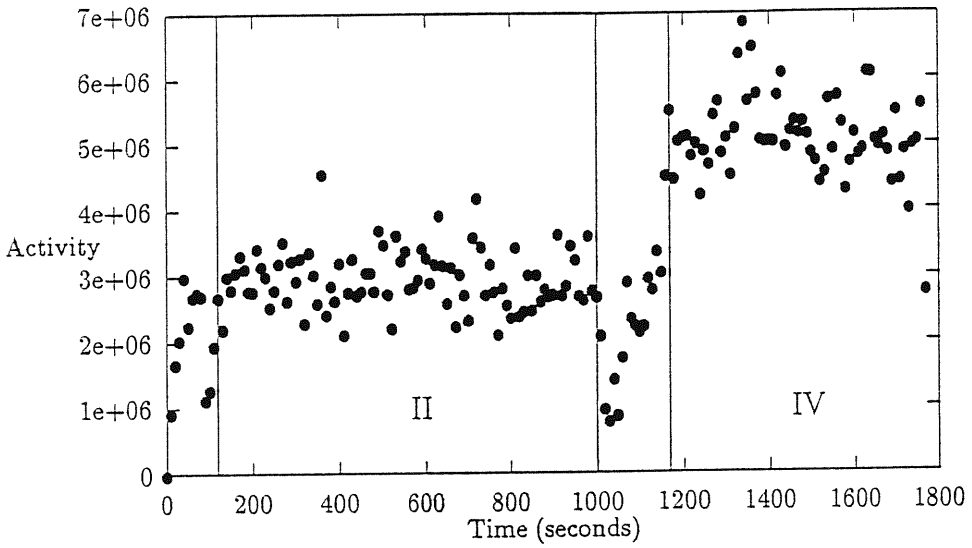

Figure 2: Activity plot for the 29 minute data set, aggregated over 10-second time intervals.

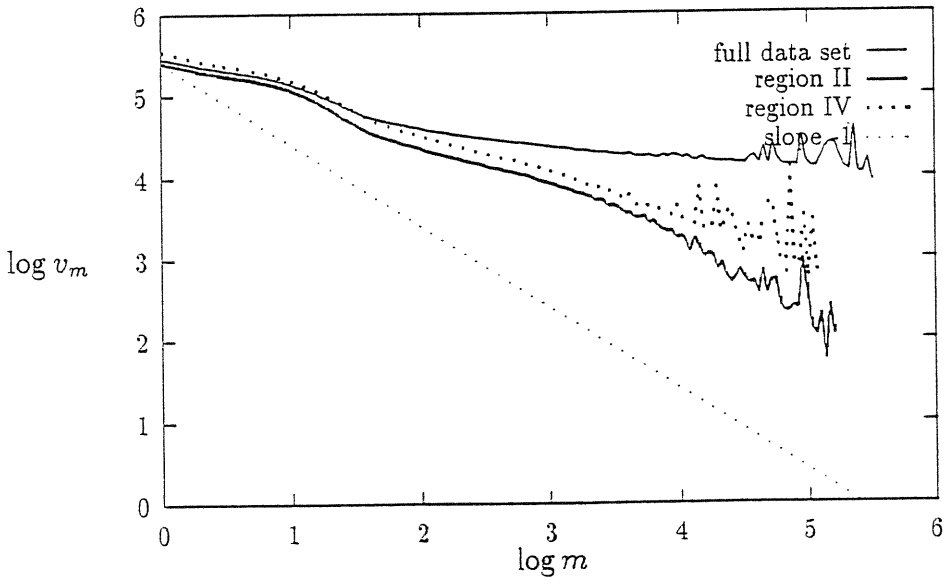

Figure 3: Variance-time plots for regions II and IV in the 29-minute data set.

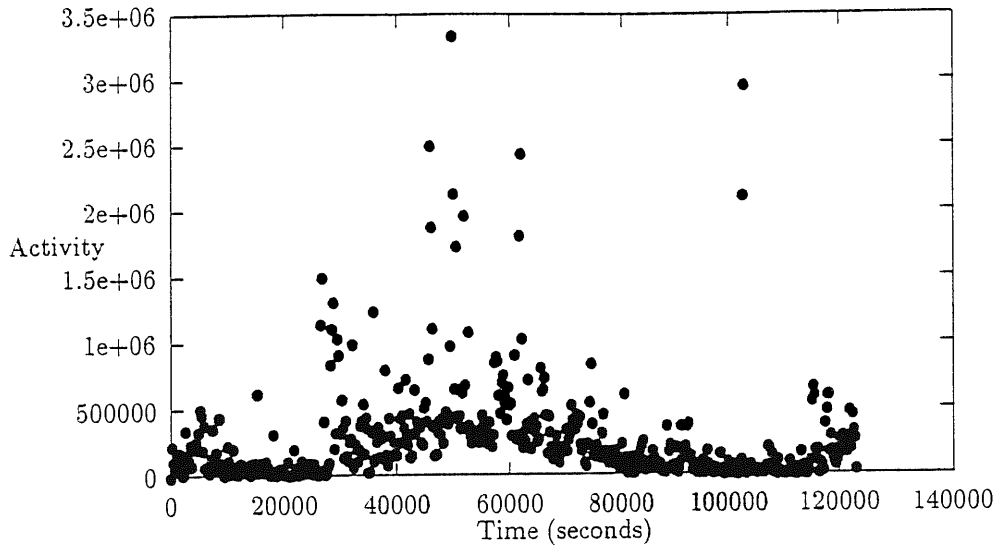

Figure 4: Activity plot for the 30-hour data set, aggregated over 240-second time intervals.

'virtual circuits' that are active and the level of burstiness displayed by each individual circuit: aggregation of independent bursty sources increases the temporal range of dependence.

We are quite confident that will be many practical situations in which (1), and hence our estimation procedures, are applicable.

The Bellcore data is certainly very challenging from the point of view of forecasting. There are two reasons for this. First, it was collected from a large network (increasing the range of dependence) and second, we have limited information on relevant details such as the number of circuits active at each point in time.

We conclude by saying that the Bellcore data set is very interesting and has raised some important questions on the problem of prediction. It has also given us the opportunity to demonstrate the age-old moral that blindfold application of 'standard' statistical tests, without looking at the data and appealing to common sense, can be misleading.

Appendix: Self-similarity, long range dependence and overflow probabilities. Suppose we have a stationary arrivals process (X_k) with $EX_1 = \mu < \infty$ and a queue with deterministic service rate $s > \mu$. A sufficient condition for (1) to hold is that the scaled cumulant generating function, defined by

$$\lambda(\theta) := \lim_{n \to \infty} n^{-1} \log E e^{\theta \sum_{k=1}^{n} X_k}, \tag{3}$$

exists, is finite in some neighbourhood of the origin and differentiable on the interior of its effective domain; then the arrivals process satisfies a large deviation principle with rate function I given by the Fenchel-Legendre transform of λ:

$$I(x) = \sup_{\theta} \{\theta x - \lambda(\theta)\}, \tag{4}$$

and δ, the asymptotic decay rate for the tails of the queue-length distribution, is given by

$$\delta = \inf_{c>0} c^{-1} I(c+s). \tag{5}$$

An alternative representation of δ is

$$\delta = \sup\{\theta : \ \lambda(\theta) \leq \theta s\}. \tag{6}$$

Variants of this result have appeared in the literature: we refer the reader to [11] for a heuristic derivation, and to the recent papers of Glynn and Whitt [10] and Duffield and O'Connell [7] for proofs under very general conditions; further bibliographic details can be found in [9].

Now suppose that the arrivals process X has the property that for large m and n,

$$\sum_{k=1}^{nm} (X_k - \mu) \ \stackrel{\mathcal{D}}{\approx} \ n^H \sum_{k=1}^{m} (X_k - \mu), \tag{7}$$

for some $H \in [1/2, 1)$. Here $\stackrel{\mathcal{D}}{\approx}$ means 'approximately equal in distribution'. Then X is said to be *asymptotically self-similar with Hurst parameter* H; if $H > 1/2$ the process exhibits long range dependence. This is the sense in which Leland *et al.* [12, 13, 14] (and others) claimed to have found evidence for self-similarity with Hurst values consistently greater than $1/2$.

If (7) holds in a suitably rigorous sense and the relevant expectations are finite, then the scaled cumulant generating function defined by

$$\lambda(\theta) := \lim_{n \to \infty} n^{-2(1-H)} \log E e^{\theta n^{1-2H} \sum_{k=1}^{n} X_k} \tag{8}$$

exists and is finite in some neighbourhood of the origin. It then follows from [7, Corollary 2.3] that, under mild regularity conditions, the tails of the corresponding queue-length distribution for a queue with constant service rate $s > \mu$ satisfy

$$\lim_{b \to \infty} b^{-2(1-H)} \log P(Q > b) = -\delta, \tag{9}$$

where

$$\delta = \inf_{c>0} c^{-2(1-H)} I(c+s), \tag{10}$$

and I is the Fenchel-Legendre transform of λ. In particular, the log-probability of overflow is asymptotically linear in buffer-size if, and only if, $H = 1/2$; otherwise the decay is polynomial and depends on the value of H.

We now describe the effect of long range dependence alone, without assuming self-similarity. Note that (8) implies

$$\lim_{n \to \infty} n^{-2H} \operatorname{var} \left(\sum_{k=1}^{n} X_k \right) \in (0, \infty); \tag{11}$$

the existence of such a power law (for some $H > 1/2$) is often treated as a definition of long range dependence for finite variance processes. A more general statement is that there exists a sequence v_n with $v_n/n \nearrow +\infty$ and that (11) holds with n^{-H} replaced by v_n^{-1}; under additional hypotheses on the asymptotic behaviour of higher order moments (see, for example, [5, pp253–]) this yields a large deviation principle for X with scaling coefficients v_n and, assuming the limit

$$g(c) := \lim_{n \to \infty} \frac{v_{n/c}}{v_n} \tag{12}$$

exists for each $c > 0$, it follows from [7, Theorems 2.1 and 2.2] that

$$\lim_{b\to\infty} v_b^{-1} \log P(Q > b) = -\delta, \tag{13}$$

where

$$\delta = \inf_{c>0} g(c)I(c+s), \tag{14}$$

and I is the Fenchel-Legendre transform of the scaled cumulant generating function

$$\lambda(\theta) := \lim_{n\to\infty} v_n^{-1} \log E e^{v_n\theta \sum_{k=1}^{n} X_k/n}. \tag{15}$$

If we thus take the existence of such a scaling sequence v_n (with the property that $v_n/n \nearrow +\infty$) as a working definition of long range dependence, we conclude that we should not expect (1) to hold in the presence of long range dependence; the actual behaviour is predicted by (13).

Acknowledgements. This research was supported by grants from EOLAS and Mentec Computer Systems Ltd, under the Higher Education-Industry Cooperation Scheme. We are grateful to Walt Willinger for making the data available and one of us (NO'C) would also like to thank David Aldous, David Brillinger, Kjell Doksum and Deb Hopkins for helpful conversations.

References

[1] David Aldous. *Probability Approximations via the Poisson Clumping Heuristic.* Applied Mathematical Sciences 77, Springer-Verlag, 1989.

[2] J. Beran, R. Sherman, M.S. Taqqu and W. Willinger. Variable-bit-rate video traffic and long range dependence. To appear in *IEEE Trans. Comm.*

[3] Cheng-Shang Chang. Stability, queue length and delay of deterministic and stochastic queueing networks. Preprint, 1993.

[4] C. Courcoubetis, G. Kesidis, A. Ridder, J. Walrand and R. Weber. Admission control and routing in ATM networks using inferences from measured buffer occupancy. To appear in *IEEE Trans. Comm.*

[5] Amir Dembo and Ofer Zeitouni. *Large Deviation Techniques and Applications.* Jones and Bartlett, Boston-London, 1993.

[6] N.G. Duffield, J.T. Lewis, Neil O'Connell, Raymond Russell and Fergal Toomey. The entropy of an arrivals process: a tool for estimating QoS parameters of ATM traffic. *Proceedings of the 11th IEE Teletraffic Symposium,* Cambridge, March 1994.

[7] N.G. Duffield and Neil O'Connell. Large deviations and overflow probabilities for the general single server queue, with applications. DIAS Technical Report No. DIAS-STP-93-30, 1993.

[8] N.G. Duffield and Neil O'Connell. Large deviations for arrivals, departures, and overflow in some queues of interacting traffic. *Proceedings of the 11th IEE Teletraffic Symposium,* Cambridge, March 1994.

[9] G. de Veciana, C. Courcoubetis and J. Walrand. Decoupling bandwidths for networks: a decomposition approach to resource management. Memorandum No. UCB/ERL M93/50, University of California.

[10] Peter W. Glynn and Ward Whitt. Logarithmic asymptotics for steady-state tail probabilities in a single-server queue. *J. Appl. Prob.*, to appear.

[11] G. Kesidis, J. Walrand and C.S. Chang. Effective bandwidths for multiclass Markov fluids and other ATM sources. Preprint, 1993.

[12] W.E. Leland, M.S. Taqqu, W. Willinger, and D.V. Wilson. Ethernet traffic is self-similar: stochastic modelling of packet traffic data. Preprint, 1993.

[13] W.E. Leland, M.S. Taqqu, W. Willinger and D.V. Wilson. Statistical analysis of high time-resolution Ethernet LAN traffic measurements. *Proceedings of INTERFACE*, 1993.

[14] W.E. Leland, M.S. Taqqu, W. Willinger, and D.V. Wilson. On the self-similar nature of ethernet traffic. Presented at SIGCOMM, 1993.